

IN THE CLAIMS:

1. (Currently amended) A data processing machine implemented method of selecting data sets for use with a predictive algorithm based on data network geographical information, comprising data processing machine implemented steps of:
 - generating a first statistical distribution of a training data set;
 - generating a second statistical distribution of a testing data set;
 - comparing the first statistical distribution and the second statistical distribution to identify a discrepancy between the first statistical distribution and the second statistical distribution with respect to data network geographical information; and
 - modifying selection of entries in one or more of the training data set and the testing data set based on the discrepancy between the first statistical distribution and the second statistical distribution; and
 - using the modified selection of entries by the predictive algorithm.
2. (Currently amended) The method of claim 1, wherein the first statistical distribution and the second statistical distribution are distributions of a number of data network links from a customer data network geographical location to a web site data network geographical location.
3. (Currently amended) The method of claim 1, wherein the first statistical distribution and the second statistical distribution are distributions of a size of a click stream for arriving at a web site data network geographical location.
4. (Currently amended) The method of claim 1, wherein comparing the first statistical distribution and the second statistical distribution includes comparing one or more of a mean, mode, and standard deviation of the first statistical distribution to one or more of a mean, mode, and standard deviation of the second statistical distribution.
5. (Currently amended) The method of claim 1, wherein the first statistical distribution and the second statistical distribution are distributions of a weighted data network geographical

distance between a customer data network geographical location and a web site data network geographical locations.

6. (Currently amended) The method of claim 1, wherein the first statistical distribution and the second statistical distribution are distributions of a weighted click stream for arriving at a web site data network geographical locations.

7. (Currently amended) The method of claim 1, wherein modifying selection of entries in one or more of the training data set and the testing data set includes generating recommendations for improving selection of entries in one or more of the training data set and the testing data set, and wherein the method of claim 1 further comprises re-generating at least one of the first statistical distribution and the second statistical distribution based upon the recommendations.

8. (Previously presented) The method of claim 1, wherein the training data set and the testing data set are selected from a customer information database comprising information with respect to customers who have purchased any of goods and services over a data network, wherein the data network geographic information pertains to geographic information of the data network.

9. (Currently amended) The method of claim 1, further comprising comparing at least one of the first statistical distribution and the second statistical distribution to a statistical distribution of a customer database to determine if the training data set and the testing data set are geographically representative of a customer population represented by the customer database.

10. (Currently amended) The method of claim 1, wherein the first statistical distribution and second statistical distribution are frequency distributions of number of data network links between a customer geographical location and one or more web site data network geographical locations, and size of a click stream for arriving at one or more web site data network geographical locations.

11. (Currently amended) The method of claim 9, wherein comparing at least one of the first statistical distribution and the second statistical distribution to a statistical distribution of a

customer database includes:

generating a composite data set from the training data set and the testing data set; and

generating a composite statistical distribution from the composite data set that was generated from the training data set and the testing data set.

12. (Currently amended) The method of claim 1, wherein modifying selection of entries in one or more of the training data set and the testing data set includes changing one of a random selection algorithm and a seed value for the random selection algorithm, and then re-comparing the first statistical distribution and the second statistical distribution.

13. (Currently amended) The method of claim 1, ~~further comprising~~ wherein using the modified selection of entries by the predictive algorithm includes training a the predictive algorithm using at least one of the training data set and the testing data set if the discrepancy is within a predetermined tolerance.

14. (Original) The method of claim 13, wherein the predictive algorithm is a discovery based data mining algorithm.

15. (Currently amended) An apparatus for selecting data sets for use with a predictive algorithm based on data network geographical information, comprising:

a statistical engine; and

a comparison engine coupled to the statistical engine, wherein the statistical engine generates a first statistical distribution of a training data set and a second distribution of a testing data set, the comparison engine compares the first statistical distribution and the second distribution to identify a discrepancy between the first statistical distribution and the second distribution with respect to data network geographical information, modifies selection of entries in one or more of the training data set and the testing data set based on the discrepancy between the first statistical distribution and the second distribution, and provides the modified selection of entries for use by the predictive algorithm; and

a predictive algorithm device that uses the modified selection of entries and the predictive algorithm.

16. (Currently amended) The apparatus of claim 15, wherein the first statistical distribution and the second statistical distribution are distributions of a number of data network links from a customer data network geographical location to a web site data network geographical location.

17. (Currently amended) The apparatus of claim 15, wherein the first statistical distribution and the second statistical distribution are distributions of a size of a click stream to arrive at a web site data network geographical location.

18. (Currently amended) The apparatus of claim 15, wherein the comparison engine compares the first statistical distribution and the second statistical distribution by comparing one or more of a mean, mode, and standard deviation of the first statistical distribution to one or more of a mean, mode, and standard deviation of the second statistical distribution.

19. (Currently amended) The apparatus of claim 15, wherein the first statistical distribution and the second statistical distribution are distributions of a weighted number of data network links between a customer data network geographical location and a web site data network geographical location.

20. (Currently amended) The apparatus of claim 15, wherein the first statistical distribution and the second statistical distribution are distributions of a weighted size of a click stream to arrive at a web site data network geographical location.

21. (Currently amended) The apparatus of claim 15, wherein the comparison engine modifies selection of entries in one or more of the training data set and the testing data set by generating recommendations for improving selection of entries in one or more of the training data set and the testing data set, and wherein the statistical engine re-generates at least one of the first statistical distribution and the second statistical distribution based upon the recommendations.

22. (Previously presented) The apparatus of claim 15, further comprising a training data set/testing data set selection device that selects the training data set and the testing data set from a

customer information database comprising information with respect to customers who have purchased any of goods and services over a data network, wherein the data network geographic information pertains to geographic information of the data network.

23. (Currently amended) The apparatus of claim 15, wherein the comparison engine further compares at least one of the first statistical distribution and the second statistical distribution to a statistical distribution of a customer database to determine if the training data set and the testing data set are geographically representative of a customer population represented by the customer database.

24. (Currently amended) The apparatus of claim 15, wherein the first statistical distribution and second statistical distribution are frequency distributions of a number of data network links between a customer data network geographical location and one or more web site data network geographical locations, and a size of a click stream to arrive at one or more web site data network geographical locations.

25. (Currently amended) The apparatus of claim 23, wherein the comparison engine compares at least one of the first statistical distribution and the second statistical distribution to a statistical distribution of a customer database by:

generating a composite data set from the training data set and the testing data set; and
generating a composite statistical distribution from the composite data set that was generated from the training data set and the testing data set.

26. (Currently amended) The apparatus of claim 15, wherein the comparison engine modifies selection of entries in one or more of the training data set and the testing data set by changing one of a random selection algorithm and a seed value for the random selection algorithm, and then re-comparing the first statistical distribution and the second statistical distribution.

27. (Currently amended) The apparatus of claim 15, ~~further comprising a predictive algorithm device~~, wherein the predictive algorithm device is trained using at least one of the training data set and the testing data set if the discrepancy is within a predetermined tolerance.

28. (Original) The apparatus of claim 27, wherein the predictive algorithm is a discovery based data mining algorithm.

29. (Currently amended) A computer program product in a computer readable medium comprising a data structure for enabling a data processing machine to select data sets for use with a predictive algorithm based on data network geographical information, comprising:

first instructions for generating a first statistical distribution of a training data set;

second instructions for generating a second statistical distribution of a testing data set;

third instructions for comparing the first statistical distribution and the second statistical distribution to identify a discrepancy between the first statistical distribution and the second statistical distribution with respect to data network geographical information; ~~and~~

fourth instructions for modifying selection of entries in one or more of the training data set and the testing data set based on the discrepancy between the first statistical distribution and the second statistical distribution; and

fifth instructions for using the modified selection of entries by the predictive algorithm.

30. (Currently amended) The computer program product of claim 29, wherein the first statistical distribution and the second statistical distribution are distributions of a number of data network links from a customer data network geographical location to a web site data network geographical location.

31. (Currently amended) The computer program product of claim 29, wherein the first statistical distribution and the second statistical distribution are distributions of a size of a click stream to arrive at a web site data network geographical location.

32. (Currently amended) The computer program product of claim 29, wherein the third instructions for comparing the first statistical distribution and the second statistical distribution include instructions for comparing one or more of a mean, mode, and standard deviation of the first statistical distribution to one or more of a mean, mode, and standard deviation of the second statistical distribution.

33. (Currently amended) The computer program product of claim 29, wherein the first statistical distribution and the second statistical distribution are distributions of a weighted number of data network links between a customer data network geographical location and a web site data network geographical location.

34. (Currently amended) The computer program product of claim 29, wherein the first statistical distribution and the second statistical distribution are distributions of a weighted size of a click stream to arrive at a web site data network geographical location.

35. (Currently amended) The computer program product of claim 29, wherein the fourth instructions for modifying selection of entries in one or more of the training data set and the testing data set include instructions for generating recommendations for improving selection of entries in one or more of the training data set and the testing data set, and wherein the computer program product claim 29 further comprises instructions for re-generating at least one of the first statistical distribution and the second statistical distribution based upon the recommendations.

36. (Currently amended) The computer program product of claim 29, further comprising fifth instructions for comparing at least one of the first statistical distribution and the second statistical distribution to a distribution of a customer database to determine if the training data set and the testing data set are geographically representative of a customer population represented by the customer database.

37. (Currently amended) The computer program product of claim 29, wherein the first statistical distribution and second statistical distribution are frequency distributions of a number of data network links between a customer data network geographical location and one or more web site data network geographical locations, and a size of a click stream to arrive at one or more web site data network geographical locations.

38. (Currently amended) The computer program product of claim 36, wherein the fifth instructions include:

instructions for generating a composite data set from the training data set and the testing

data set; and

instructions for generating a composite distribution from the composite data set that was generated from the training data set and the testing data set.

39. (Currently amended) The computer program product of claim 29, wherein the fourth instructions for modifying selection of entries in one or more of the training data set and the testing data set include instructions for changing one of a random selection algorithm and a seed value for the random selection algorithm, and then re-comparing the first statistical distribution and the second statistical distribution.

40. (Currently amended) The computer program product of claim 29, ~~further comprising~~ wherein the fifth instructions include instructions for training a the predictive algorithm using at least one of the training data set and the testing data set if the discrepancy is within a predetermined tolerance.

41. (Previously presented) A data processing machine implemented method of predicting customer behavior based on data network geographical influences, comprising data processing machine implemented steps of:

- obtaining data network geographical information regarding a plurality of customers;
- training a predictive algorithm using the data network geographical information; and
- using the predictive algorithm to predict customer behavior based on the data network geographical information.

42. (Previously presented) An apparatus for predicting customer behavior based on data network geographical influences, comprising:

- means for obtaining data network geographical information regarding a plurality of customers;

- means for training a predictive algorithm using the data network geographical information; and

- means for using the predictive algorithm to predict customer behavior based on the data network geographical information.

43. (Previously presented) A computer program product in a computer readable medium comprising a data structure for enabling a data processing machine to predict customer behavior based on data network geographical influences, comprising:

first instructions for obtaining data network geographical information regarding a plurality of customers;

second instructions for training a predictive algorithm using the data network geographical information; and

third instructions for using the predictive algorithm to predict customer behavior based on the data network geographical information.